PYSPARK DEMO

BENI SHPRINGER

DS6003

MOTIVATION

- Use PySpark to:
  - Save a data set using into a parquet format
  - Read from that parquet format into a PySpark dataframe
  - Manipulate the dataframe to prepare for machine learning techniques
  - Apply MLlib functions to perform machine learning on data
  - Visualize data and glean insights from the MLlib analysis
- Use a baseball data set to see if K rate and BB rate affect BABIP
  - Many contend that BABIP is a "random" statistic that is not a function of other parts of a players performance – I wanted to see if this is true by comparing it to K rate and BB rate.

CODE SNIPPET

```
from pyspark.ml.linalg import Vectors, VectorUDT
from pyspark.ml.feature import VectorAssembler
assembler = VectorAssembler(
    inputCols=["BB_pct", "K_pct", "SB"],
    outputCol="features")
```

```
train_output = assembler.transform(trainingDF)
test_output = assembler.transform(testDF)
```

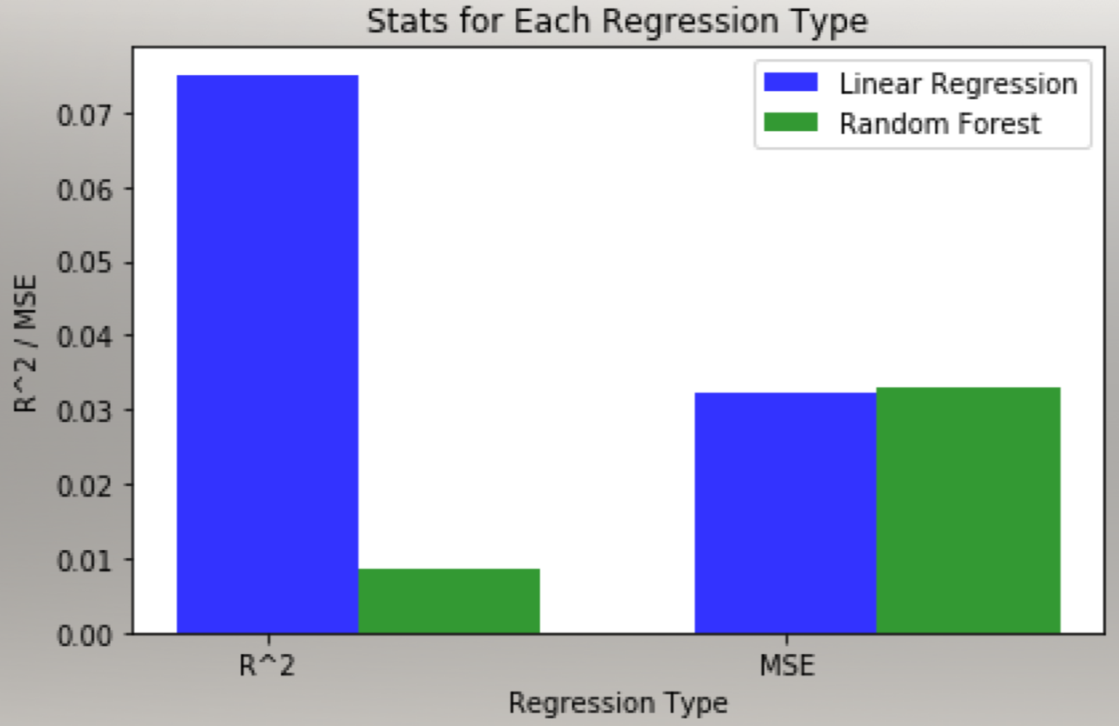
```
test_output.take(1)
```

```
[Row(BB_pct=0.049, K_pct=0.201, BABIP=0.31, SB=24, features=DenseVector([
0.049, 0.201, 24.0]))]
```

```
trainingDF_renamed = train_output.withColumnRenamed("BABIP", "label").selec
testDF_renamed = test_output.withColumnRenamed("BABIP", "label").select('fe
```

- Here is where I simply figured out how to add multiple features for my models using the VectorAssembler. This took me a while to figure out, but once I did, I was able to achieve what I wanted by feeding all the features into the model.
- I'll know what to do more quickly next time I want to do ML with PySpark.

VISUALIZATION



- The left shows some statistics from the testing run using both linear regression and random forest models. The two models had similar MSE, but the Linear Regression Model had a better R^2.
- Regardless, this analysis shows that BABIP is a very difficult statistic to predict. I thought using three statistics which are not related to hitting the ball might provide some insights, but alas, they were not predictive of this stat which many believe is sheer luck.